AD______________

AWARD NUMBER DAMD17-97-1-7193

TITLE: Methods for Evaluating Mammography Imaging Techniques

PRINCIPAL INVESTIGATOR: Carolyn M. Rutter, Ph.D.

CONTRACTING ORGANIZATION: Group Health Cooperative of Puget Sound
Seattle, Washington 98101-1448

REPORT DATE: June 1998

TYPE OF REPORT: Annual

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE June 1998	3. REPORT TYPE AND DATES COVERED Annual (19 May 97 - 18 May 98)

4. TITLE AND SUBTITLE

Methods for Evaluating Mammography Imaging Techniques

5. FUNDING NUMBERS

DAMD17-97-1-7193

6. AUTHOR(S)

Carolyn M. Rutter, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Group Health Cooperative of Puget Sound
Seattle, Washington 98101-1448

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research And Materiel Command
ATTN: MCMR-RMI-S
504 Scott Street
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Dr. Rutter sought this Department of Defense Breast Cancer Research Program Career Development Award to further develop the skills, knowledge, and experience that will enable her to develop biostatistical methods for breast cancer research. She is on track with her stated goals. During the first award year, she has succeeded in gaining knowledge about breast cancer epidemiology, and has made significant progress toward statistical research goals. Evaluation of a bootstrap approach to accuracy estimation is nearly complete. This evaluation extends the research proposed in the original application to include a comparison against a new estimator, a description of a bootstrap estimate that adjusts for verification bias, and a description of the bootstrap estimator in a multi-reader study.

14. SUBJECT TERMS **receiver operating characteristic curve, diagnostic accuracy**
Breast Cancer

15. NUMBER OF PAGES
25

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

X For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

6-18-98

PI - Signature Date

Table of Contents

I. Introduction

Dr. Rutter sought this Department of Defense Breast Cancer Research Program Career Development Award to further develop the skills, knowledge, and experience that will enable her to develop biostatistical methods for breast cancer research. Her original statement of work is provided in Appendix A. During the first award year, she has succeeded in gaining knowledge about breast cancer epidemiology, and has made significant progress toward statistical research goals.

Dr. Rutter's statistical research program focuses on receiver operating characteristic (ROC) curves for ordinal test outcomes. ROC curves analysis is a method of describing and comparing diagnostic test performance when test outcomes are ordinal or continuous. ROC curves are a natural extension of analysis via sensitivity and specificity. Sensitivity is the probability that a test correctly identifies a disease-positive case. Specificity is the probability of correctly identifying a disease-negative case. ROC curves model the trade-off between sensitivity and specificity as the criterion for a positive test is varied. ROC terminology makes the trade-off between sensitivity and specificity explicit: "true positive rate" (TP) replaces sensitivity and "false positive rate" (FP) replaces 1-specificity. As the proportion of positive tests increases, both TP and FP tend to increase. An ordinal test with k possible outcomes generates up to k-1 different (FP, TP) pairs, or operating points. Each operating point results from choosing a different cutpoint for determining a positive test. Two implicit operating points represent extreme test behavior (0,0), when all tests are negative; and (1,1), when all tests are positive. The ROC curve is constructed by plotting and linking these k+1 operating points. The area under the ROC curve (AUC) provides a single number summary of overall test performance. An uninformative test has an equal chance of a positive result among diseased-positive and disease-negative cases for every criterion, corresponding to an ROC curve that is a 45 line and an AUC statistics equal to 1/2. The AUC provides a simple method for comparing the performance of competing tests.

During the first award year, Dr. Rutter focused on aim 1 of her statistical research goals: developing methods for estimating accuracy statistics (TP, FP, AUC) when patients are assessed at multiple patient sites. She has finished a review of current research for generalized estimating equation and random effect approaches for nonlinear models, and is finalizing an article describing bootstrap estimation of accuracy statistics (TP, FP, AUC) when patients are assessed at multiple patient sites (Appendix B).

VIII. Achievement of Year 1 Technical Objectives

Technical Objective 1: Gain additional training in breast cancer epidemiology, detection and treatment.

Dr. Rutter has developed an understanding of breast cancer that will guide her development of statistical methods. Directed readings, guided by Drs. Margaret Mandelson and Emily White, provided her with basic information about breast cancer etiology, incidence, progression and diagnosis. A key reference was a 1993 Epidemiologic Review of breast cancer.[1] During this first award year, the Seattle Breast Cancer Research Program discontinued its seminars. However, throughout the 1997/1998 academic year Dr. Rutter has participated in a Diagnostic Methods working group, lead by Dr. Mary Lou Thompson, a research associate professor at the University of Washington's Department of Biostatistics. Topics addresssed by this groups include regression models for estimating diagnostic accuracy and treatment of errors in the gold standard for true disease state. Dr. Rutter has also attended Breast Cancer Surveillance group meetings. Surveillance meetings have provided her with important practical information about radiologists' interpretation of mammograms, and the timing and execution of diagnostic procedures.

Dr. Rutter's understanding of the etiology and progression of breast cancer will underlie the statistical methods she develops for analyzing mamographic accuracy. Breast cancer results from mutations that occur during epithelial cell proliferation in the ducts and lobes. Hormones, such as estrogen and progesterone, affect breast cancer risk through their effect on cell proliferation rates. Breast cancer represents one extreme on a continuum of disease, ranging from benign proliferative disease, to carcinoma in situ, and finally infiltrating carcinoma.[2] Disease definitions also need to allow for the occurrence of more than one pathological type within a single lesion. The definition of a breast cancer case is an important aspect of study design since the definition of breast cancer affects the apparent accuracy of mammography.

Breast cancer research also confirms the need to develop methods that account for error in the gold standard. Because mammography is a screening tool, false negative mammograms (also called 'interval cancers') are identified using follow-up information. Longer follow-up periods allow more complete capture of false negative cases, but risk inclusion of incident cancers and omission of women who are lost to follow-up. Furthermore, because disease can be present at multiple diffuse foci, and because pathology can vary within a lesion it is possible for biopsy to miss disease that is present.

The method of collecting of mammographic interpretation data also affects statistical models. The standardized set of mammographic interpretations proscribed by the American College of Radiology lexicon improves data collection by virtue of standardizaton.[3] However, the inclusion of an interpretive code for additional work-up complicates evaluation of mammographic accuracy. The additional work-up category does not fit neatly into an ordinal outcome scale. These cases include a mix of women, for example, it could naturally include both women with suspected cysts (benign disease) and women with suspicious findings that need additional evaluation. Models need to be developed to handle these kinds of data. One possible approach to these data is extension of two-part models employed in econometrics.[4] The first part of the model would estimate the probability of an interpretation based on the current mammogram (i.e., additional workup not requested). The second part of the model would describe ordinal outcomes among observations with an interpretation of the current mammogram. Inference is drawn from the combined results from these two model steps.

Technical Objective 2: Develop methods for multiple patient assessments.

During the past year, new research has significantly advanced statistical methodology for receiver operating characteristic (ROC) curves. Leisenring and Pepe describe generalized estimating equation (GEE) approaches to diagnostic test assessment.[5,6,7] The ROC models proposed by Pepe have significantly advanced statistical methodology.[7] These models can accommodate correlated rating data, and estimation of models can be carried out using standard statistical software packages.

Recent work by Walsh demonstrates limitations of the robustness of ROC curve analysis.[8]. Prior to Walsh's work, the explicit assumption was that parametric ROC curve estimation for ordinal test outcomes was insensitive to the latent variable model.[9] This assumption was based on both simulation results and heuristic arguments. However, Walsh shows that bias in the estimated area under the ROC curve increases as the maximum observed false positive rate decreases. Thus, comparisons between tests can be misleading when there is wide discrepancy in maximum observed false positive rates. In light of the development of nonparametric GEE models for ROC analysis and description of the sensitivity of parametric ROC model to latent distribution assumptions, Dr. Rutter has abandoned refinements to simple parametric approaches, such as random effects modeling and robust covariance adjustment to account for correlated data.

Dr. Rutter has continued work describing, evaluating, and applying nonparametric bootstrap ROC estimation for correlated data. The most recent draft of this article is provided in Appendix B. This scope of this research has been expanded and now incorporates comparisons against a newly developed alternative approach proposed by Obuchowski.[10] Obuchowski describes a non-iterative method for nonparametrically estimating the area under the ROC curve (AUC) that uses sums of squares to adjust variance estimates for correlation between observations. During the last year, Dr. Rutter completed simulations comparing Obuchowski's estimator to the bootstrap estimator. Both methods are theoretically valid, and both perform well in a simple situation. However, the bootstrap estimator can be used in more complex sampling situations that include multiple sources of correlation.

Dr. Rutter has also extended her evaluation of the bootstrap estimator to include estimation in the face of verification bias. Verification bias occurs when the probability that a patient's disease status is verified with a gold standard assessment depends on the studied test. For example, this occurs when patients with 'clearly negative' tests are not sent to surgery. Verification bias is common, and there are well-known adjustments for verification bias.[11] Bootstrap estimates can be weighted to account for verification bias, and these estimates incorporate the additional variability of bias-adjusting weights. Simulations confirming the consistency of verification-bias adjusted bootstrap estimates are near completion.

Imaging tests are often evaluated by more than one reader. The frequency of multi-reader tests has increased during recent years, with increasing awareness of between reader variability.[12] The bootstrap approach allows estimation of overall test performance based on individual reader outcomes. Two recent articles use this approach. Halpern and colleagues compare computed tomography and ultrasound assessments of renal artery stenosis.[13] This study had two readers independently evaluate images from each modality. Every film was evaluated twice. Readers provided two assessments per patient, one for each renal artery. The bootstrap approach was used to compare readers, and to compare average reader accuracy for ultrasound and computed tomography. The bootstrap estimate was also used to analyze data from IMAGE (Improving Mammography Accuracy with Group Evaluation) substudy of the Breast Cancer Surveillance. This analysis described the accuracy of 31 mammographers who participated in this rereading study. Each mammographer evaluated the same set of 140 screening mammograms and provided disease ratings for each breast. Simulations confirming the consistency of multi-reader bootstrap estimates are currently underway.

Technical Objective 3: Develop and teach a course in methods for assessing diagnostic tests.

During the past year Dr. Rutter gave an invited presentation, "Meta-Analysis of Diagnostic Test Data", at the 1997 International Conference on Health Policy Research held December 5-7 at Crystal City, Virginia. She also presented this research at the Diagnostic Methods working group. This research will be incorporated into a lecture for the diagnostic methods course.

III. Progress Toward Other Grant Aims

Collaboration with investigators participating in the Breast Cancer Surveillance Consortium has highlighted the need for methods that allow for error in gold standard information. This is particularly important for evaluation of breast cancer. Because mammography is used primarily as a screening tool, the gold standard is generally constructed by combining information from biopsy results and mammographic follow-up. Several authors have explored methods for estimating test accuracy when there are multiple test outcomes with no true gold standard.[14-17] Some articles have described methods that allow estimation of accuracy in the absence of gold standard information.[18,19] Interest in this area is increasing as researchers begin to face the inherent uncertainty of a 'definitive' diagnosis.

IV. Summary

Dr. Rutter is on track with her stated goals. She has made significant progress towards proposed research goals. Specific tasks and objectives for the first award year were:

1. Gain additional training in breast cancer epidemiology, detection and treatment.
   a. Review of information on the epidemiology, diagnosis and treatment of breast cancer as suggested by Dr. Margaret Mandelson.
   b. Attend seminars sponsored by the Seattle Breast Cancer Research Program. (through year 4)
2. Statistical research, aim 1: develop methods for multiple patient assessments
   a. Review current research for generalized estimating equation and random effect approaches for nonlinear models.
   b. Test bootstrap, robust covariance adjustment and generalized estimating equation methods for breast-level analyses using simulation studies.
3. Develop and teach a course in methods for assessing diagnostic tests
   a. Collect relevant references and outlining lectures for the methods course. During this time, specific lectures may be presented in other University of Washington courses (through year 2).

Evaluation of the bootstrap approach is nearly complete. This evaluation extends the research proposed in the original application to include a comparison against a new estimator, a description of a bootstrap estimate that adjusts for verification bias, and a description of the bootstrap estimator in a multi-reader study.

IV. References

1. Kelsey JL, Ed. *Epidemiologic Reviews: Breast Cancer, Volume 15, No. 1,* The Johns Hopkins University School of Hygiene and Public Health: Baltimore, Maryland, 1993.
2. Bodain CA. "Benign Breast Diseases, Carcinoma In Situ, and Breast Cancer Risk," *Epidemiologic Reviews,* 15: 177-187, 1993.
3. Linver MN, Osuch JR, Brenner RJ, Smith RA. "The Mammography Audit: A Primer for the Mammography Quality Standards Act (MQSA)," *American Journal of Roentgenology,165:19-25, 1995.*
4. Judge GG, Griffiths WT, Hill RC, Lütkepohl H, Lee T. *The Theory and Practice of Econometrics, Second Edition,* John Wiley and Sons: New York, 1980.
5. Leisenring W, Pepe MS, Longton G. "A Marginal Regression Modelling Framework for Evaluating Medical Diagnstotic Tests," *Statistics in Medicine,* 16: 1263-1281, 1997.
6. Pepe MS. "A Regression Modelling Framework for ROC Curves in Medical Diagnostic Testing," *Biometrka,* in press.
7. Pepe MS. "Interpretation, Estimation and Regression for ROC Curves," draft article, presented at the April 1998 Surveillance Consortium Meeting, Seattle WA.
8. Walsh SJ. "Limitations to the Robustness of Binormal Roc Curves: Effects of Model Misspecification and Location of Decision Thresholds on Bias, Precision, Size and Power," *Statistics in Medicine,* 16: 669-679, 1997.
9. Hanley JA. "Receiver Operating Characteristic (ROC) Methodology: The State of the Art," *Critical Reviews in Diagnostic Imaging,*29:307-335, 1989.
10. Obuchowski NA. "Nonparametric Analysis of Clustered ROC Curve Data," *Biometrics,* 53: 567-578, 1997.
11. Gray R, Begg CB, Greenes RA. "Construction of Receiver Operating Characteristic Curves when Disease Verification is Subject to Selection Bias," *Medical Decision Making,* 4: 151-164, 1984.
12. Obuchowski NA, Zepp RC. "Simple Steps for Improving Multiple-Reader Studies in Radiology," *American Journal of Roentgenology,* 166: 517-521, 1996.
13. Halpern EJ, Rutter C, Gardiner GA, Nazarian LN, Wechsler RJ, Outwater EK, Mitchell DG, Kueny-Beck M, Levin DB, Moritz MJ, Carabasi RA, Kahn MB, Smullens SN, Feldman HI. "Comparison of Doppler Ultrasound, CT Angiography and MR Angiography for Evaluation of Renal Artery Stenosis," *Academic Radiology,* in press.
14. Walter SD, Irwig LM. "Estimation of Test Error Rates, Disease Prevalence and Relative Risk from Misclassified Data: A Review," *Journal of Clinical Epidemiology,*41: 923-937, 1988.
15. Epstein LD, Muñoz A, He D. "Bayesian Imputation of Predicitve Values When Covariate Information is Available and Gold Standard Diagnosis is Unavailable," *Statistics in Medicine,* 15: 463-476, 1996.
16. Lu Y, Keying Y, Mathur AK, Hui S, Feurst TP, Genant HK. "Comparative Calibration Without a Gold Standard," *Statistics in Medicine,* 16: 1889-1905, 1997.
17. Torrance- Rynard VL, Walter SD. "Effects of Dependent Errors in the Assessment of Diagnostic Test Performance," *Statistics in Medicine,*16: 2157-2175, 1997.
18. Joseph L, Gyorkos TW, Coupal L. "Bayesian Estimation of Disease Prevalence and the Parameters of Diagnostic Tests in the Absence of a Gold Standard," *American Journal of Epidemiology,*141: 263-272, 1995.
19. Joseph L, Gyorkos TW. "Inferences for Likelihood Ratios in the Absence of a 'Gold Standard'," *Medical Decision Making,* 16: 412-417, 1996.

VII. Appendices

Appendix A. Statement of Work

Technical Objective 1: Gain additional training in breast cancer epidemiology, detection and treatment.

Task 1: Months 1-4: Review of information on the epidemiology, diagnosis and treatment of breast cancer as suggested by Dr. Margaret Mandelson.

Task 2: Months 1-48: Attend seminars sponsored by the Seattle Breast Cancer Research Program.

Technical Objective 2: Statistical research, aim 1: develop methods for multiple patient assessments.

Task 3: Month 6: Review current research for generalized estimating equation and random effect approaches for nonlinear models.

Task 4: Months -11: Test bootstrap, robust covariance adjustment and generalized estimating equation methods for breast-level analyses using simulation studies.

Task 5: Months 12-21: Develop methods for woman-level analysis, possibly including software development for random effects in generalized ordinal regression models.

Technical Objective 3: Statistical research, aim 2: extend exact methods for ordinal regression models

Task 6: Month 22: Review current research in exact methods.

Task 7: Months 23-34: Extend exact methods and write computational algorithms and programs to compute distributions of sufficient statistics.

Technical Objective 4: Statistical research, aim 3: Develop methods to adjusting for measurement error in disease status

Task 8: Month 36: Review current research in errors-in-measurement models.

Task 9: Months 37-48: Develop simple combined corrections for verification and follow-up bias. These methods will be extended to allow adjustments in general ordinal regression models.

Technical Objective 5: Develop and teach a course in methods for assessing diagnostic tests.

Task 10: Months 1-24: Collect relevant references and outlining lectures for the methods course. During this time, specific lectures may be presented in other University of Washington courses.

Task 11: Months 25-36: Offer methods course at University of Washington through the Department of Biostatistics.

APPENDIX B: Bootstrap estimation of diagnostic accuracy using patient-clustered data

Carolyn M. Rutter
Center for Health Studies, Group Health Cooperative
1730 Minor Avenue, suite 1600, Seattle, WA 98101

Abstract

Ordinal outcomes are common in radiology research, with radiologists rating their confidence that disease is present (or absent) based on diagnostic images. Correlated data arise naturally in this setting. When comparing diagnostic modalities, the efficiency of comparisons is increased by assessing patients with each tested modality. When evaluting the location of disease, patients are assessed at multiple body sites. Using information from multiple body sites can also increase the efficiency of comparisons between tests. Recently developed methods allow estimation and comparisons of diagnostic accuracy when multiple body sites are assessed within patients. This paper describes bootstrap estimation of true positive rates, false positive rates, and the area under the reciever operating characteristic curve (AUC) when multiple sites are evaluated within patients. Bootstrap estimates are consistent because these accuracy statistics are generalized U-statistics. The bootstrap approach can be used to describe the accuracy of a single test and to compare the accuracy two or more tests. Bootstrap estimation is easy to apply, even in complicated sampling settings. In a simple setting, we found that the size and coverage rates of the bootstrap AUC estimator were very similar to the size and coverage rates of Obuchowski's estimator based on estimation of cross-products.[3] We also demonstrate bootstrap estimation in a more complicated setting: using data from a mammographer rereading study we estimate mammographers' overall accuracy and individual mammographers deviation from the group average. In this setting boostrap estimates incorporated correlation resulting from both the mammographers' evaluation of the same set of films from 113 patients and evaluation of up to 2 per patient.

keywords: receiver operating characteristic curve, true positive rate, false positive rate, simulation.

1. Introduction

The diagnostic accuracy of imaging tests is often estimated from correlated ordinal outcomes. This type of data is common in radiology reasearch, where radiologists are asked to evaluate diagnostic imaging studies and provide interpretations using an ordinal scale. Correlated diagnostic test outcomes arise because of evaluation by multiple tests and because of evalution of multiple body sites. Evaluation of patients using multiple tests usually occurs by design. Because of between patient variability, comparisons between diagnostic tests are more efficient when each patient is evaluated using all compared tests Not surprisingly, the first methods developed for correlated test data addressed correlation arising from multiple assessment of a single site.[?, 2] In this setting, each patient either has disease or does not have disease. Evaluation of multiple sites within patients often occurs because of practical application of diagnostic tests. Examples of multi-site diagnostic assessments include screening mammography to detect breast cancer,[4] computed tomography of the liver to detect metastatic colorectal cancer,[5] and magnetic resonance angiography of leg vessels to detect occlusive peripheral vascular disease.[6] In each of these cases, disease can be located at a particular site (or sites) and correct localization is required for surgical treatment. An important feature of multi-site assessments is that disease state can vary between sites.

Statistical methods for combining correlated test results from multiple sites are relatively new. Two methods have recently been proposed. Obuchowsky describes a method for estimating standard errors for the area under the empirical receiver operating characteristic curve area under the curve (AUC) based on sums of squares.[3] Obuchowsky's method allows estimation of the standard error of the AUC for a single test, or the standard error of the difference between AUC statistics for two tests. Estimation of the appropriate sums of squares can become more complicated when there are additional sources of correlation. One way this can occur is when each test was evaluated by more than one reader. Pepe proposes a general regression methodology that allows comparison between two tests, and uses bootstrap resampling to estimate standard errors.[7] This regression approach can account for multiple sources of correlation. Furthermore, regression methodology allows inclusion of covariates, including continuous covariates. However, because the regression approach uses new statistical methods, estimation of true positive rates, false positive rates, and AUC statistic are useful when comparing new results to previously published finding We propose a simple bootstrap approach for estimating true positive rates, false positive rates, and AUC for patient-clustered data. The bootstrap approach is especially useful when there are multiple sources of correlation.

2. Nonparametric measures of test accuracy: *tp*, *fp*, AUC

When test outcomes are dichotomous, true positive rates and false positive rates measure test accuracy. A test's true positive rate (*tp*) estimates the probability of a positive outcome (indicating presence of disease) when the target disease is present. Similarly, a test's false positive rate (*fp*) estimates the probability of a positive test outcome when disease is absent. A perfect diagnostic test has $fp = 0$ and $tp = 1$. When test outcomes are ordinal, *tp* and *fp* can be calculated by dichotomizing outcomes. However, a single (fp, tp) pair cannot completely describe the accuracy of an ordinal test because both *tp* and *fp* rates depend on test stringency.

Reciever operating characteristic (ROC) curve analysis accounts for the tradeoff between *tp* and *fp* as test stringency varies. Suppose the ordinal outcome of a diagnostic test, t_i, takes values in $\{1, 2, \ldots, K\}$ with increasing values of t_i corresponding to stronger evidence of disease. There are $K + 1$ possible ways to dichotomize the ordinal test, including 'all positive' and 'none positive', and each is associated with a (tp, fp) pair. The empirical ROC curve is drawn by plotting pairs of observed rates, fp versus *tp*, and connecting the $K + 1$ consecutive points with straight lines. The empirical ROC curve provides a simple graphical description of test performance.

The overall accuracy of an ordinal test can be summarized by the area under the ROC curve (AUC). The AUC estimates the probability of correctly ranking a randomly selected (diseased,not-disease) pair on the ordinal test scale; It ranges from 0 to 1, with the value 1 corresponding to a perfect diagnostic test. A test that is no better than chance has an AUC equal to one half. The AUC is asymptotically normally distributed. The test of H_o : AUC $= 1/2$ based on the asymptotic distribution is equivalent to a Mann-Whitney test. This test based on the AUC is a test for differences in the distribution of ordinal test outcomes in diseased versus not-diseased groups.[8]

Each of these nonparametric accuracy statistics (*tp*, *fp*, AUC) is a generalized U-statistic: Each one is a sum of functions of statistically independent quantities.[9] Because *tp*, *fp*, and

AUC are U-statistics, bootstrap resampling provides consistent point and interval estimates.[11]

Bootstrap samples are constructed by drawing patients, the independent units, with replacement. This incorporates all sources of within patient variability. To ensure that accuracy statistics are estimable, bootstrap samples are stratified by patient-level disease state. This corresponds to conditioning on true disease state. Accuracy statistics are calculated for each bootstrap sample. The accuracy of two tests can be compared by calculating the difference in accuracy statistics for each bootstrap sample. This incorporates between test correlation. Point estimates are simple averages of statistics, or the differences between in statistics, across bootstrap samples. Standard errors are estimated using the observed standard errors across bootstrap samples. Standard errors should be based on at least 100 draws. Confidence intervals are estimated using bootstrap estimated standard errors, with a normal approximation. Confidence intervals can also be estimated using percentiles, though this requires at least 1,000 samples.[12]

When data are correlated, the U-statisitic properties of estimates must be maintained to ensure consistency of bootstrap estimates. In other words, *tp*, *fp*, and AUC must each remain a sum of functions of statistically independent quantities. Application of the bootstrap to *tp* and *fp* rates is straightforward. Suppose each patient is evaluated at up to m sites. Let $\mathbf{t_i} = (t_{i1}, t_{i2}, \ldots, t_{im})'$, be the vector of test outcomes across these m sites, and let $\mathbf{d_i} = (d_{i1}, d_{i2}, \ldots, d_{im})'$, be the corresponding vector of 0/1 disease indicators. True positive and false positive rates for the k^{th} cutpoint are:

$$tp_k = \frac{1}{n_D}\sum_i \phi_k(\mathbf{t_i}, \mathbf{d_i}) \qquad \text{and} \qquad fp_k = \frac{1}{n_{\overline{D}}}\sum_i \phi_k(\mathbf{t_i}, (\mathbf{1} - \mathbf{d_i}))$$

with kernel function $\phi_k(\mathbf{t_i}, \mathbf{d_i}) = \sum_j \delta_k(t_{ij})d_{ij}$ where $\delta_k(t) = 1$ if $t \geq k$, and is otherwise zero. The associated sample sizes are $n_D = \sum_i \sum_j d_{ij}$ and $n_{\overline{D}} = \sum_i \sum_j (1 - d_{ij})$. Here D indicates presence of disease and $\overline{D}$ indicates absence of disease. Even when data are correlated, *tp* and *fp* rates are sums of independent quantities, so that calculations of bootstrap estimates is straightforward.

Maintaining the U-statistic properties of the AUC statistic while resampling clustered data requires some care. The empirical AUC statistic is:

$$\text{AUC} = \frac{\sum_{i \in D} \sum_{j \in \overline{D}} \psi(t_i, t_j)}{n_D n_{\overline{D}}}$$

with kernel function

$$\psi(t_i, t_j) = \begin{cases} 1 & \text{if} \quad t_i > t_j \\ \frac{1}{2} & \text{if} \quad t_i = t_j \\ 0 & \text{if} \quad t_i < t_j \end{cases}$$

These sums are over diseased (D) and not diseased ($\overline{D}$) groups. When patients have both diseased and not-diseased sites, ratings from D and $\overline{D}$ groups are correlated and the empirical AUC statistics is not a generalized U-statistic. The independence of D and $\overline{D}$ groups can be maintained by excluding correlated $(D, \overline{D})$ pairs from calculations. This approach excludes direct comparisons within patients. The new kernel function is:

$$\psi_m(t_{ij}, t_{i'j'}) = \begin{cases} 1 & \text{if} \quad t_{ij} > t_{i'j'} \text{ and } i \neq i' \\ \frac{1}{2} & \text{if} \quad t_{ij} = t_{i'j'} \text{ and } i \neq i' \\ 0 & \text{if} \quad t_{ij} < t_{i'j'} \text{ or } i = j' \end{cases}$$

the sum, $\Sigma_{ij \in D} \Sigma_{i'j' \in \overline{D}} \psi_m$, is divided by the total number of independent ratings that contribute information. The AUC statistic based on the kernel function ψ_m estimates the probability of correctly ranking an independent (D,$\overline{D}$) pair.

Bootstrap estimation is especially useful when data are clustered and variance formulae are not available, for example when sites are clustered within patients and verification of true state depends on test outcome. Let V indicate verification status, with $V = 1$ if disease status is verified by a gold standard assessment, and $V = 0$ if not verified. The probability of a test outcome, T, given disease state, D can be caculated as a function of observed probabilities using Bayes rule:[14]

$$P(T = t|D = d) = \frac{P(D = d)P(D = d|V = 1)}{P(T = t)P(T = t|V = 1)} P(T = t, D = d, V = 1)$$

where $P(D = d) = \sum_{t=1}^{K} P(T = t)P(D = d|T = t, V = 1)$. When verification across sites within patients is independent, accuracy estimates can be corrected for verification bias by weighting verified observations using:

$$w(t, d) = \frac{\widehat{P}(T = t)}{\widehat{P}(T = t|V = 1)} \frac{\widehat{P}(D = d|V = 1)}{\widehat{P}(D = d)}$$

Test outcomes from unverified patients contribute to accuracy estimates through estimated weights. More complicated weights that account for different verification scenarios are possible. Bias corrected accuracy statistics are generalized U-statistics with unknown parameters.

Bias-corrected tp and fp rates are:

$$tp_k^{bc} = \frac{1}{n_D} \sum_{i:d_i=1} \phi_k^{bc}(t_i, d_i, v_i) \quad \text{and} \quad fp_k^{bc} = \frac{1}{n_{\overline{D}}} \sum_{i:d_i=0} \phi_k^{bc}(t_i, d_i, v_i)$$

with kernel function $\phi_k^{bc}(\mathbf{t_i}, \mathbf{d_i}) = \sum_j \delta_k(t_{ij}) d_{ij} w(t_{ij}, d_{ij}, v_{ij})$ where, as before, $\delta_k(t) = 1$ if $t \geq k$, and is otherwise zero. The associated sample sizes are $n_D = \sum_i \sum_j d_{ij} w(t_{ij}, d_{ij}, v_{ij})$ and $n_{\overline{D}} = \sum_i \sum_j (1 - d_{ij}) w(t_{ij}, d_{ij}, v_{ij})$.

The kernel function for the adjusted empirical ROC curve is:

$$\psi_m^{bc}(t_{ij}, t_{i'j'}, w_{ij}, w_{i'j'}) = \begin{cases} w(t_{ij}, 1) w(t_{i'j'}, 0) & \text{if} \quad t_{ij} > t_{i'j'} \text{ and } i \neq i' \\ \frac{1}{2} w(t_{ij}, 1) w(t_{i'j'}, 0) & \text{if} \quad t_{ij} = t_{i'j'} \text{ and } i \neq i' \\ 0 & \text{if} \quad t_{ij} < t_{i'j'} \text{ or } i = j' \end{cases}$$

The $2K$ distinct weights used to adjust for verification bias are estimated parameters. The denominator for the AUC statistic is the weighted sum of independent comparisons. The bias-corrected version of AUC, is a generalized U-statistic with $2K$ unknown parameters, so that it is asymptotically normally distributed and bootstrap resampling methods provide consistent point and interval estimates.[16]

3. Small sample behavior: simulation study design & results

Using a brief simulation study, we compared bootstrap AUC estimates to Obuchowski's AUC estimates. Comparisons are based on the size and power of tests for differences between two AUC statistics based on the normal approximation, and on coverage rates for a single AUC statistic. Simulated data represent outcomes from two tests, for 4 sites within 100 patients. We assume that half the patients have disease at one or more sites. For each patient with at least on disease-positive site, the number of additional affected sites is simulated using a binomial random number generator with probability 0.5. Test A, the standard test, has an empirical AUC equal to 0.8 and false positive rates equal to (0.05, 0.1, 0.3, 0.5). Test B, the new test, hast the same false positive rates. The AUC for test B is equal to 0.8 or 0.9. Each test has a 5-point ordinal outcome.

Ordinal test outcomes were simulated by categorizing continuous multivariate normal (MVN) psuedodeviates. One standard MVN pseudodeviate of length $2m$ was generated for each patient-observation using the IMSL subroutine DRMVN[13]. (Recall that m is the number of sites within patients.) Cutpoints on the ordinal scale were created using fixed false positive rates. The four cutpoints are $\theta_k = \Phi^{-1}(1 - fp_k)$. Given these cutpoints, the MVN mean of the disease positive sites was shifted so that the rates produced the desired AUC. The mean shift, μ, was estimated iteratively for each AUC value. True positive rates were calculated from the shifted distribution, $tp_k = \Phi(\theta_k + \mu)$, and the underlying AUC was calculated from (tp, fp) points. μ was updated until the calucated AUC was within 0.00001 of the desired AUC.

We explored four types of within patient correlation: none, moderate, high within-patient correlation, and high between-test correlation. Within patient correlation was estimated on the MVN scale. Let t_{ijk} be the rating for the outcome of the k^{th} test at the j^{th} site within the i^{th} patient, and let D_{ijk} be the corresponding true disease state. Then

$$\operatorname{corr}(t_{ijk}, t_{i'j'k'}) = \begin{cases} \rho_{P0} & \text{if} \quad i = i' \quad D_{ijk} = D_{i'j'k'} = 0 \quad k = k' \\ \rho_{P1} & \text{if} \quad i = i' \quad D_{ijk} = D_{i'j'k'} = 1 \quad k = k' \\ \rho_{P2} & \text{if} \quad i = i' \quad D_{ijk} \neq D_{i'j'k'} \quad k = k' \\ \rho_{T1} & \text{if} \quad i = i' \quad j = j' \quad k \neq k' \\ \rho_{T2} & \text{if} \quad i = i' \quad j \neq j' \quad k \neq k' \\ 0 & \quad \text{otherwise} \end{cases}$$

This allows the correlation between sites within patients to depend on the true state of these sites. If $\rho_{P0} = \rho_{P1} = \rho_{P2}$ then some patients tend to have high scores overall, while others tend to have low scores. If tests results are less variable when disease is present, then ρ_{P1} is is greater than both ρ_{P0} and ρ_{P2}. The correlation structures we examined are given in Table 1.

Characteristics of the bootstrap sample when data are subject to verification bias were examined by setting disease rates to missing with probability 0.4 when the standard test was given one of the two lowest ratings.

Characteristics of the bootstrap sample when two readers assess provide ratings for each patient were examined by generating a vector of length $4m$ for each patient, and assuming that $2m$ were readings from the first reader, and the other $2m$ were from the second reader. As before, rating data were generated by categorizing multivariate normal data to produce desired false positive rates and true AUC statistics. Between reader correlation was generated on the continuous scale. We assume that separate readers were used for the two tests, so that between reader correlation for different tests were set to zero. Within test reader correlation was set to 0.25. We also assume that within tests the readers are equally accurate, with a true AUC equal to either 0.8 or 0.9.

Simulation Results:

Obuchowski's noniterative estimator and the bootstrap estimator had very similar test characteristics (Table 2). Both methods had size and coverage near the nominal levels. The power to detect a difference between AUC_A and AUC_B was also similar for the two methods. Results show expected trends for the power to detect a difference in AUC statistics. Power was lowest when there was high within patient corelation, but relatively low between test correlation. Power was highest when there was high between test correlation.

The weighted bootstrap estimator had good performance when data were subject to verification bias (Table 3). When verification bias is ignored, estimates are optimistic, with size greater than 5% and coverage rates greater than 95%.

4. Mammographer Rereading Study

The greatest benefit of bootstrap estimation is seen when correlation structures are relatively complicated. In this example we use the bootstrap estimator to measure individual and overall diagnostic accuracy of 31 mammographers who took part in a rereading study. Each mammographer evaluated the same set of screening mammograms from 113 women and gave separate interpretations for each breast. These interpretations were given using a five point scale: 1) negative or benign; 2) probably benign (short interval follow-up needed); 3) possibly abnormal (additional views needed); 4) suspicious abnormality (biopsy should be considered); 5) highly suggestive of malignancy. The film set included mammograms from 30 women who had pathology-verified unilateral breast cancer and mammograms from 87 women who did not have breast cancer. True disease state was based on a combination of information from biopsy and two years of follow-up mammography.[4]. Bootstrap estimation was used to estimate the overall performance of the group of mammographers, and each individual mammographer's deviation from the group's average performance.

Bootstrap samples were created by drawing patients with replacement. For each bootstrap sample we calculated four types of oucomes: 1) individual mammographers' four possible (tp, fp) pairs (operating points); 2) individual mammographers' AUC statistics (AUC_i); 3) the average AUC across mammographers ($\overline{\text{AUC}}$); and 4) the difference between each mammographers' AUC and the average across the remaining mammographers' AUC: $\mathrm{AUC}_i - \overline{\mathrm{AUC}_{(i)}}$. This approach accounts for correlation resulting from assessment of two sites per patient and from multiple assessments of the same films.

Before reporting averaged measures, we checked their appropriateness by examining individual mammographer's empirical ROC curves, using their bootstrap estimated operating points. Individual mammographers' ROC curves were qualitatively similar. Across mammographers, bootstrap AUC estimates ranged from 0.83 to 0.93. The overall mean, $\overline{\mathrm{AUC}}$, was 0.88 with 95% confidence interval (0.80,0.93). Based on estmated deviations from the remaining mammographers, no mammographer was significantly different from the others. We concluded that in this group practice setting the overall mean, $\overline{\mathrm{AUC}}$, accurately represents the screening accuracy of this group of mammographers.

5. Discussion

Diagnostic evaluation often involves testing at multiple sites within patients. Recent methodological developments allow simple comparisons of correlated AUC statistics. In particular, we found that Obuchowski's noniterative method for calculating the AUC had good properties. The bootstrapping AUC estimator had similar properties. We use simulation studies to compar Obuchowski's method to the bootstrap method in several other scenarios, varying the numbers of sites, the sample prevalence, the true AUC statistics, and true false positive rates. In each case our findings were qualitatively similar: There were no differences in the performance of these two approaches. The similarity between these approaches is reassuring, since both approaches are theoretically valid.

The primary advantage of the bootstrap estimator is that it easily generalizes to complex sampling designs. For example, in the mammography rereading study correlation results from evaluation of two breasts per woman and from repeated assessment of films across mammographes. In this case, the bootstrap approach allowed estimation of overall average AUC statistics, and each mammographers' deviation from this overall mean. Before calculating an overall AUC statistic, individual ROC curves need to be examined. Combined results are sensible only when individual ROC curves have similar shapes. Similarly, the ROC curve from each test should be examined before before comparing two tests using AUC statistics. When ROC curves have very different shapes, focusing on the area under these curves can mask important differences.

Another advantage of bootstrap estimation is that bootstrap estimators naturally incorporate all available data. Data must be missing at random for valid inference. However, when missing data depends on test outcomes, bootstrap estimators can be adjusted for verification bias[14] by applying the adjustment to each bootstrap sample. These bias-adjusted AUC statistics are consistent estimators, since they are simply U-statistics with unknown parameters.[15, 16]

The bootstrap approach presented in this article maintained U-statistic properties of the AUC estimator. This allowed reliance on U-statistic theory and ensured consistency of estimates. Anecdotally, a simple bootstrap that included comparisons between correlated $(D, \overline{D})$ pairs performed just as well. However, such good performance may not hold in general.[17]

The greatest limitation of AUC statistics is their inability to incorporate covariate information. Stratification can be used to explore covariate effects on the AUC, though this approach breaks down as either the number of covariates or the number of covariate levels increases. When covariate effects are important, regression models, such as those proposed by Pepe[7], are appropriate. In this context, bootstrap accuracy estimates can serve an important function, allowing description of covariate effects in terms of standard accuracy statistics.

References

[1] Metz CE, Wang P, Kronman HB. "A new approach for testing the signficance of differences betwen ROC curves measured from correlated data," in *Information Processing in Medical Imaging,* Deconinck F. (ed), Nijhof: The Hague, 1984.

[2] Toledano AY, Gatsonis CG. "Ordinal Regression Methodology for ROC Curves Derived from Correlated Data," *Statistics in Medicine,* 15: 1807-1826, 1996.

[3] Obuchowski NA. "Nonparametric Analysis of Clustered ROC Curve Data" *Biometrics,* 53: 567-578.

[4] Taplin SH, Rutter CM, Elmore J. "Mammographic screening and screening-diagnostic skills among radiologists: variability within and correlations between" *submitted to JAMA,* April, 1998.

[5] Zerhouni EA, Rutter CM, Hamilton SR, Balf DM, Megibow AJ, Francis IR, Moss AA, Heiken JP, Tempany CMC, Aisen AM, Weinreb J, Gatsonis C, McNeil BJ. "CT and MRI imaging in the staging of colorectal carcinoma: Report of the Radiologic Diagnostic Oncology Group II," *Radiology,* 200: 443–51, 1996.

[6] Baum RA, Rutter CM, Sunshine JH, Blebea JS, Blebea J, Carpenter JP, Dickey KW, Quinn SF, Gomes AS, Grist TM, McNeil BJ for the American College of Radiology Rapid Technology Assessment Group. "Multi-center trial to evaluate peripheral vascular magnetic resonance angiography," *Journal of the American Medical Association,* 274: 875-880, 1995.

[7] Pepe MS. "Three approaches to regression analysis of receiver operating characteristic curves for continuous test results," *Biometrics,* 54: 124-135, 1998.

[8] Hanley JA and McNeil BJ. "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* 143: 29–36, 1982.

[9] Lee AJ. *U-Statistics, Theory and Practice,* New York: Marcel Decker, 1990.

[10] Bickel PJ, Freedman DA. "Some asymptotic theory for the bootstrap," *The Annals of Statistics,* 9: 1196–1217, 1981.

[11] Arcones MA, Giné E. "On the bootstrap of U and V statistics," *The Annals of Statistics,* 20: 655–674, 1992.

[12] DiCiccio TJ, Efron B. (1996) "Bootstrap Confidence Intervals," *Statistical Science,* 11:189–212.

[13] International Mathematical and Statistical Libraries *IMSL Stat Library Users Manual (Version 1.1),* IMSL Inc: Sugarland, Texas, 1989/

[14] Gray R, Begg CB, Greenes RA. "Construction of receiver operating characteristic curves when disease verification is subject to selection bias," *Medical Decision Making,* 4: 151–164, 1984.

[15] Randles RH. "On the asymptotic normality of statistics with estimated parameters," *The Annals of Statistics,* 10: 462–474, 1982.

[16] Janssen P, Veraverbeke N. "Bootstrapping U-Statistics with Estimated Parameters," *Communications in Statistics,* 21: 1585–1603; 1992.

[17] Young GA. "Bootstrap: More Than a Stab in the Dark?" *Statistical Science,* 9: 382–415; 1994.

Table 1: Correlation Structure

none	$\rho_{P0} = \rho_{P1} = \rho_{P2} = \rho_{T1} = \rho_{T2} = 0$
moderate	$\rho_{P0} = \rho_{P1} = \rho_{P2} = \rho_{T1} = \rho_{T2} = 0.25$
high within patient	$\rho_{P0} = \rho_{P2} = 0.5$, $\rho_{P1} = 0.75$, $\rho_{T1} = \rho_{T2} = 0.25$
high between test	$\rho_{P0} = \rho_{P1} = \rho_{P2} = \rho_{T2} = 0.25$, $\rho_{T1} = 0.75$

Table 2: Observed test performance, based on 1,000 simulations with AUC_A=0.8

		size	power	coverage
correlation structure	estimator	$AUC_B = 0.8$	$AUC_B = 0.9$	$AUC_B = 0.9$
none	obuchowski	5.1	91.6	93.3
	bootstrap	5.3	91.7	93.9
moderate correlation	obuchowski	5.0	96.3	94.9
	bootstrap	4.7	95.6	94.9
high within patient correlation	obuchowski	6.9	83.7	93.3
	bootstrap	6.4	82.8	93.8
high between test correlation	obuchowski	6.2	100.0	93.5
	bootstrap	5.6	100.0	94.3

Table 3: Observed performance of bootstrap in the presence of verification bias, based on 1,000 simulations with AUC_A=0.8

	size	power	coverage
correlation structure	$AUC_B = 0.8$	$AUC_B = 0.9$	$AUC_B = 0.9$
none	5.7	88.2	93.9
moderate correlation	7.4	91.7	95.0
high within patient correlation	6.6		
high between test correlation			

* remaining simulations are underway

Table 4: Observed performance of bootstrap when two readers assess each film, based on 1,000 simulations with AUC_A=0.8

	size	power	coverage
correlation structure	$AUC_B = 0.8$	$AUC_B = 0.9$	$AUC_B = 0.9$
none			
moderate correlation			
high within patient correlation			
high between test correlation			